NPS55-78-008

NAVAL POSTGRADUATE SCHOOL

Monterey, California

SOME STATISTICAL PROCEDURES FOR THE JOINT OIL ANALYSIS PROGRAM

by

D. R. Barr

T. Jayachandran

H. J. Larson

May 1978

Prepared for:
AP-TSC NARF/Code 360
nsacola, Fla. 32508

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA

Rear Admiral T. F. Dedman
Superintendent

J. R. Borsting
Provost

Reproduction of all or part of this report is authorized.

This report was prepared by:

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NPS55-78-008	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)*: Some Statistical Procedures for the Joint Oil Analysis Program	5. TYPE OF REPORT & PERIOD COVERED: Technical
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): D. R. Barr, T. Jayachandran, H. J. Larson	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Postgraduate School, Monterey, Ca. 93940	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: P.O. # MME-77-006
11. CONTROLLING OFFICE NAME AND ADDRESS: JOAP-TSC NARF/Code 360, Pensacola, Fla. 32508	12. REPORT DATE: March 1978
	13. NUMBER OF PAGES: 72
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*	15. SECURITY CLASS. *(of this report)*: Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Oil Analysis
Standard Artification
Laboratory Certification
Electrode Acceptance Tests

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Procedures are described for

1. Acceptance testing of prepared oil standards.
2. Certification of spectrometric laboratories.
3. Acceptance testing of graphite electrodes for use in the oil analysis program.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601 |

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

SOME STATISTICAL PROCEDURES FOR THE JOINT OIL ANALYSIS PROGRAM

FINAL REPORT FOR PROJECT ORDER MME-77-006

by

D. R. Barr

T. Jayachandran

H. J. Larson

Naval Postgraduate School
Monterey, CA 93940

May 1978

TABLE OF CONTENTS

SOME STATISTICAL PROCEDURES FOR THE JOINT OIL ANALYSIS PROGRAM

FINAL REPORT FOR PROJECT ORDER MME-77-006

by

D. R. Barr, H. J. Larson and T. Jayachandran

I. INTRODUCTION

The Joint Oil Analysis Program is a tri-service standardized program to monitor equipment wear condition through the use of oil analysis. Spectrometric oil analysis is used to determine the type and amount of wear metals in lubricating fluid samples. There are three primary factors that can affect the accuracy and effectiveness of oil analysis.

1. The daily spectrometer calibration routine and the particular oil standard used in the calibration.
2. The electrode type used in the analysis.
3. The experience and training of the spectrometer operator/evaluator.

This report describes statistical procedures developed under a project sponsored by the Joint Oil Analysis Program Technical Support Center, Pensacola, Florida and funded by the Engineering Division, Kelly Air Force Base, San Antonio, Texas.

Statistical procedures for acceptance testing of new batches of calibration standards are described in Section II. A three-part statistical procedure for certification of the spectrometric laboratories is presented in Section III. Section IV deals with statistical acceptance tests of electrodes from different suppliers.

In all three sections certain results of analyses of experimental data supplied by the TSC are quoted. These data consisted of acceptance testing readings of prepared oil standards by three laboratories under ideal conditions. Since these ideal conditions are not expected to occur in routine daily work, one should be careful not to extrapolate these results to more general situations. The numbers used in the worked examples came from the same source and, again, may not be typical of what can be expected in day-to-day laboratory work. The authors would like to acknowledge the kind and generous assistance of Mr. Richard S. Lee, Senior Army Representative of the Joint Oil Analysis Program Technical Support Center, Pensacola, Florida. Any errors of reasoning which may remain are the sole responsibility of the authors.

II. CALIBRATION STANDARDS

II.1. Introduction

The methods and criteria we suggest for acceptance testing of Calibration Standards are an adaptation of accepted statistical procedures, to accommodate specific features of JOAP data. We therefore begin with a discussion of some features of these data, based on sampling the calibration data provided us by the JOAP-TSC. Next, the problem of determining tolerance values (both for accuracy and repeatability) is discussed, with reference to the Baird Atomic acceptance numbers and the tolerances published by the JOAP-TSC. Finally, a test procedure is suggested for determining acceptability of new reference standards.

II.2. Characteristics of JOAP data

Various data sets of the calibration test data provided by JOAP-TSC were sampled, to provide estimates of variance-covariance matrices as well as Repeatability Index characteristics over elements, laboratories and concentrations. As an example, we show in Table 1 estimated variances (on the main diagonal), covariances (above the main diagonal) and correlations

	FE	AG	AL	CR	CU	MG	NA	NI	PB	SI	SN	TI	MO
FE	7.78	3.44	6.00	1.67	2.22	5.00	2.89	2.44	13.11	11.00	4.89	7.22	10.11
AG	.48	6.62	2.98	.73	2.58	1.76	1.33	2.02	7.29	6.04	3.40	5.71	3.02
AL	.87	.47	6.10	2.58	2.68	4.43	4.94	2.57	12.14	10.02	4.20	4.63	5.46
CR	.42	.20	.74	2.01	1.52	2.14	3.39	1.41	4.12	3.82	1.62	.88	-.70
CU	.58	.73	.79	.78	1.88	1.46	2.50	1.66	4.97	4.31	2.13	2.03	.43
MG	.57	.22	.57	.48	.34	9.88	2.72	1.57	10.14	8.58	3.87	5.97	5.68
NA	.38	.19	.73	.87	.66	.31	7.61	2.06	8.39	7.33	2.56	1.83	-.28
NI	.68	.61	.81	.77	.94	.39	.58	1.66	4.74	3.98	2.13	1.92	.99
PB	.92	.55	.96	.57	.71	.63	.59	.72	26.23	21.16	9.07	11.88	15.30
SI	.92	.55	.95	.63	.73	.64	.62	.72	.96	18.40	7.60	10.73	12.42
SN	.91	.68	.88	.59	.81	.64	.48	.86	.92	.92	3.73	5.04	5.24
TI	.79	.68	.57	.19	.45	.58	.20	.46	.71	.77	.80	10.68	11.03
MO	.80	.26	.49	-.11	.07	.40	-.02	.17	.66	.64	.60	.74	20.54

TABLE 1. Variance-covariance structure estimates with R-1 at 100 ppm, Corpus Christi Lab (first run in data set 5). Numbers below main diagonal are estimated correlations.

(below the main diagonal) for R-1 at 100 ppm at the Corpus Christi Lab. Typically, most of the correlations are positive and many of the correlations are quite large. For example, the estimated correlation between Pb and Aℓ analyses is .96. This means that, within a single analysis, a Pb reading above 100 was very likely to be accompanied by an Aℓ reading also above 100; indeed, the relationship between Pb in a given analysis and Aℓ in that same analysis was essentially linear (with positive slope).

Such correlations substantially complicate the computational difficulty of using a reference testing procedure that simultaneously incorporates data from all elements. Therefore, we recommend a procedure that continues the present practice of performing separate analyses with each element. Even so, the correlation among analyses for various elements (within a sample run) makes precise evaluation of overall error rates of a testing procedure difficult, a point we shall return to below.

In order to get an idea of the consistency of the repeatability index over elements, labs and time, the variance in analyses for individual sample runs was estimated for a number of situations. For example, Table 2 shows estimates made from data sets 1 and 5 in the data provided by the JOAP-TSC. From these analyses, the following conclusions were reached:

	LAB	Index	Fe	Ag	A	Cr	Cu	Mg	Na	Ni	Pb	Si	Sn	Ti	Mo
DATA SET 1	C	AI	.03	.04	.00	.23	.08	.19	.22	.24	.06	.03	.08	.31	.02
		RI	.283	.232	.365	.189	.326	.277	.346	.313	.448	.400	.365	.223	.316
	M	AI	.09	.36	.21	.44	.01	.07	.50	.05	.34	.24	.16	.03	.68
		RI	.325	.232	.465	.665	.074	.306	.330	.414	.707	.381	.905	.295	.744
	P	AI	.04	.33	.43	.32	.07	.10	.08	.11	.04	.24	.06	.19	.22
		RI	.313	.067	.467	.305	.106	.082	.123	.370	.540	.201	.611	.631	.922
DATA SET 5	C	AI	2.1	1.7	2.4	2.0	0.7	2.8	1.2	2.9	7.3	3.4	0.9	2.8	1.9
		RI	3.38	3.74	3.41	3.13	2.31	5.81	4.24	1.79	5.89	5.32	2.33	4.10	5.36
	M	AI	0.7	6.8	1.7	2.5	0.7	2.5	2.1	2.8	7.5	0.8	4.5	4.2	4.4
		RI	3.71	7.47	4.24	2.84	2.95	5.76	5.53	3.65	3.41	3.82	3.50	3.71	6.80
	P	AI	0.1	3.4	0.8	1.7	2.4	1.5	1.7	0.8	5.4	1.6	1.6	0.4	4.4
		RI	3.54	2.84	3.36	2.50	2.46	4.06	2.21	3.68	3.66	2.37	3.17	3.50	4.55
Mean RI @ 3ppm			.307	.177	.432	.386	.168	.221	.266	.366	.565	.327	.627	.383	.661
Mean RI @100ppm			3.54	4.68	3.67	2.822	2.57	5.21	3.99	3.04	4.32	3.83	3.00	3.77	5.57
RI Rank @ 3ppm			5	2	10	9	1	3	4	8	11	6	12	7	13
RI Rank @100ppm			5	11	6	2	1	12	9	4	10	8	3	7	13

TABLE 2. Accuracy Index and Repeatability Index for Data Sets 1 and 5.

1) Variances among elements may differ significantly.

2) There is a weak but discernable pattern of variance sizes among elements (for example, Cu is among the lowest and Mo is among the highest).

3) There seems to be no consistent pattern of variance sizes among labs.

4) Variance patterns among the two standards within a sample run tend to be consistent. That is, high variance in R-1 Pb tends to go with high variance in R-2 Pb for a given sample run.

5) High variance for one element in a sample run does not imply other elements in that sample run are also outside reasonable variance standards.

The above conclusions pertain to the particular data set on which they are based and may not be typical of day-to-day routine readings.

Based on these conclusions, the following recommendations are made concerning the test procedure:

1) Do the standards acceptability test separately for each element (further supporting the present procedure in this regard).

2) Since the reference standards are prepared by the TSC, and a spectrometer is available to the TSC at Pensacola, complete all reference standard acceptance testing at the TSC.

II.3. Tolerance Specifications

The data provided by the JOAP-TSC were used to investigate how the repeatability index responded to changes in concentration in a particular element, and to determine whether the response characteristics were the same for all elements. This is important since a statistical procedure will measure significance of apparent differences in mean concentration in terms of underlying repeatability of analyses. It was found that, for most elements with concentrations in the range 0-100 ppm, the repeatability index increased as quadratic functions of initial concentrations (see Figure 1). However, adequate fit for practical purposes is obtained with a linear function (that is, for practical purposes, one may assume $RI = mC_0 + b$, where C_0 is the initial concentration, m is the rate of increase in RI with C_0 and b is the intercept). As an example, Table 3 shows estimates of b and m for both the linear fit and quadratic fit. These are based on R-1 analyses at the Pensacola Lab (last run), at concentrations of 3, 10, 30, 50 and 100 ppm. Figure 2 shows plots of the linear fit for 13 elements.

It was found the elements appear to have different patterns of increase of RI with C_0. This suggests a different tolerance criterion should be used for RI for each element. Adequacy of the linear and quadratic fits are indicated by the estimated correlation values r in Table 3. Values of .95 or

Figure 1. Plots of $\sqrt{RI} = MC_0 + b$ with R-1 at five levels of C_0.

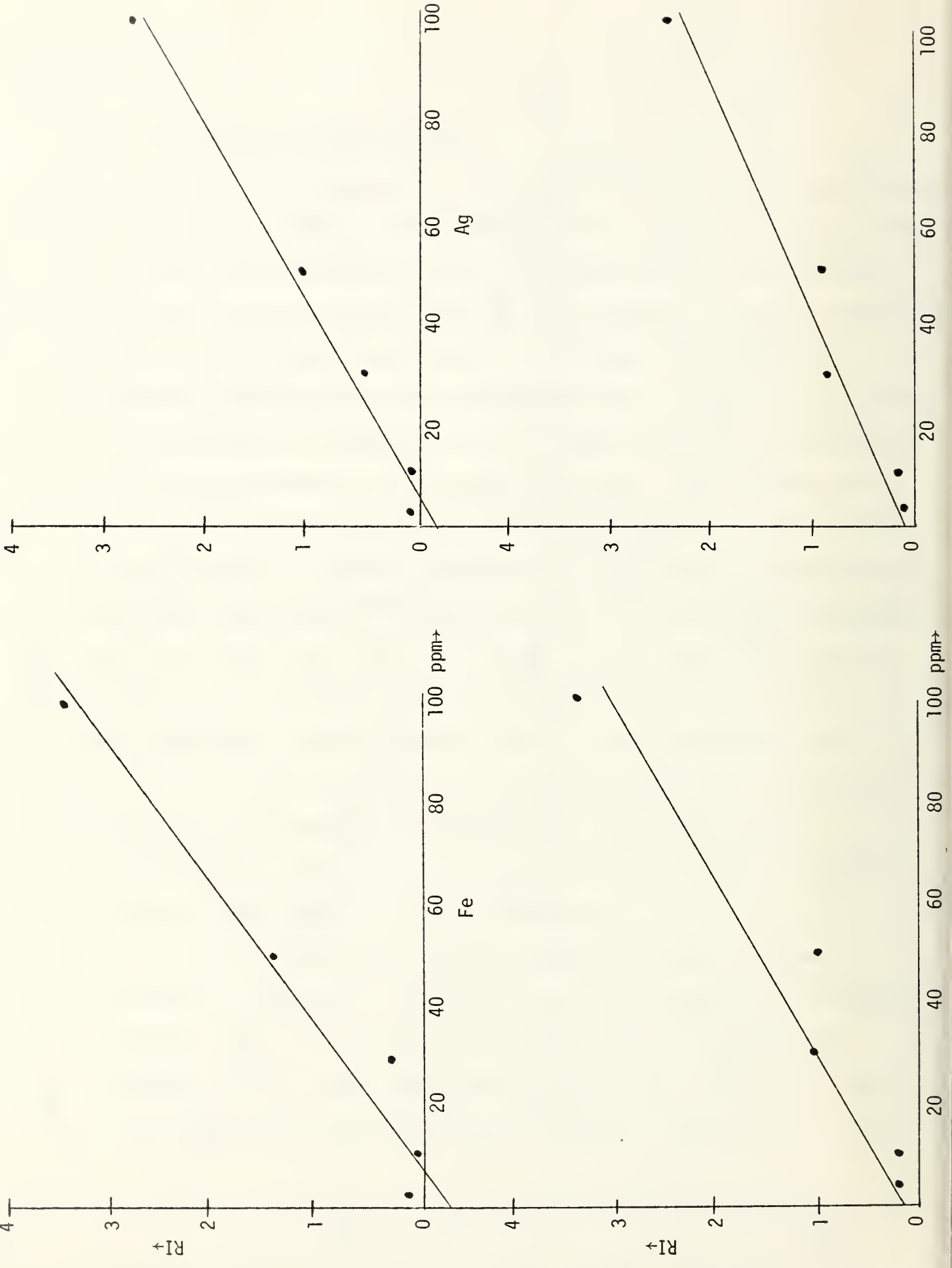
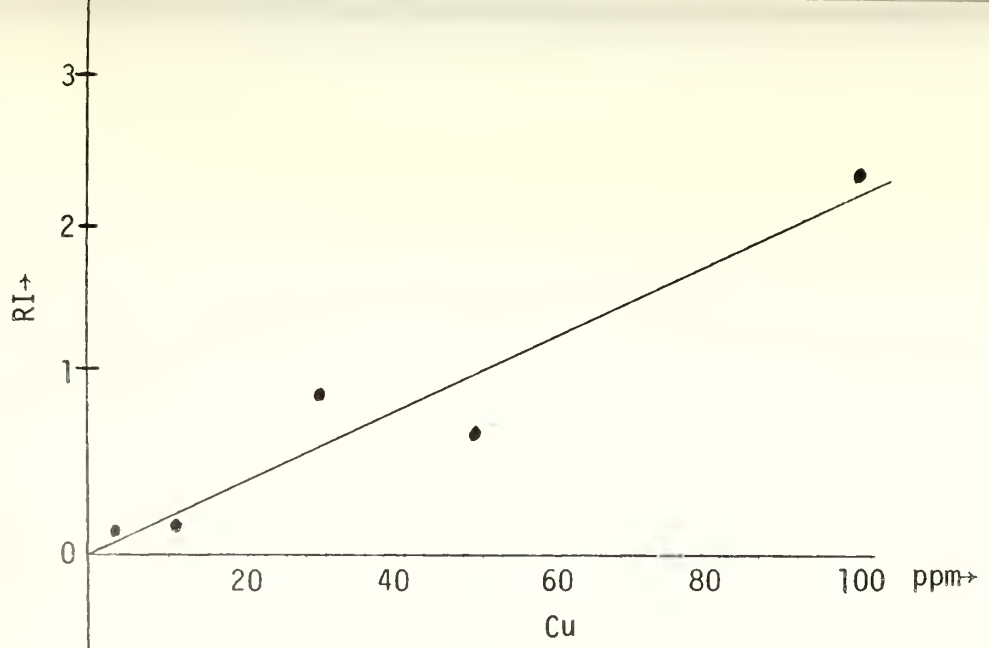
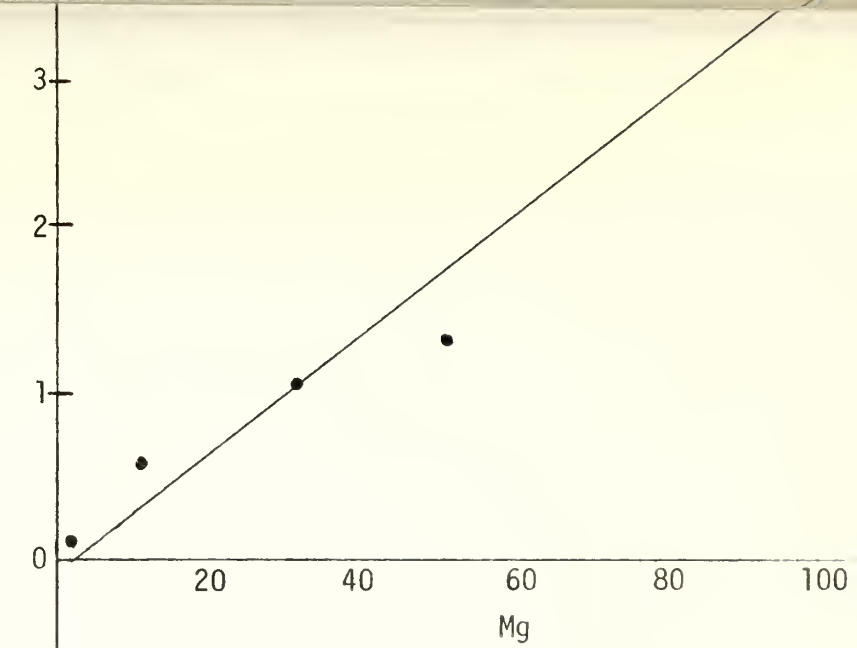
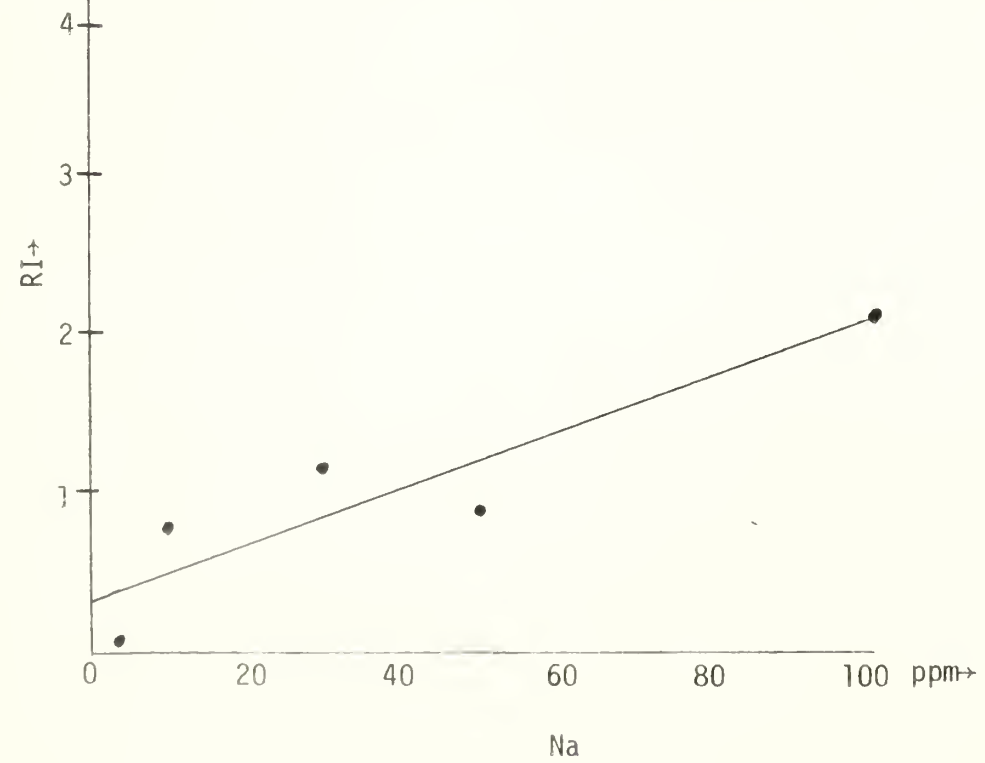


Figure 2. Plots of $RI = MC_0 + b$ with R-1 at five levels of C_0.







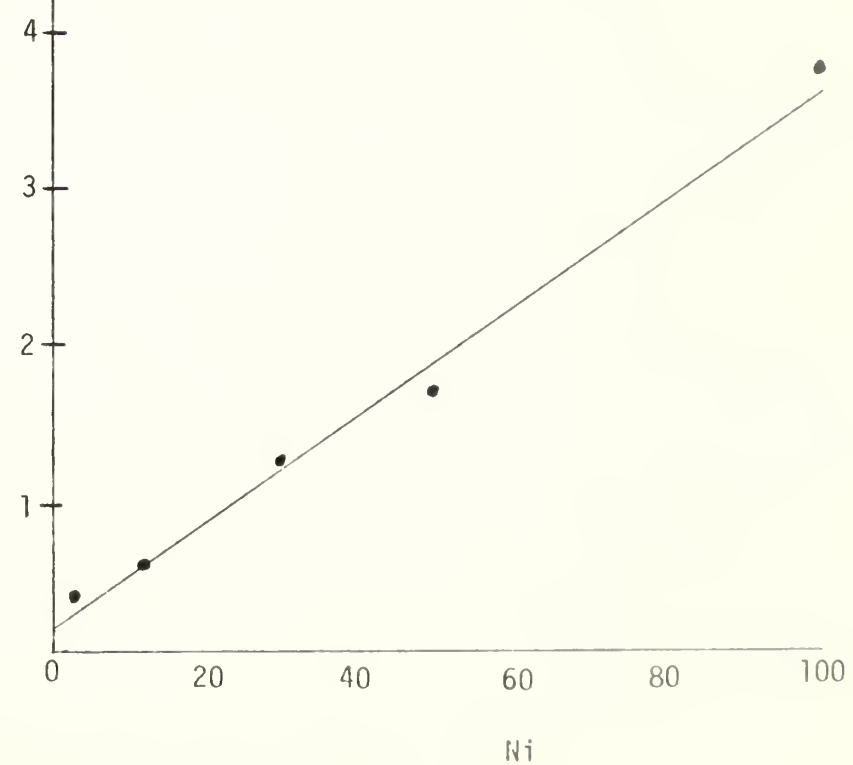
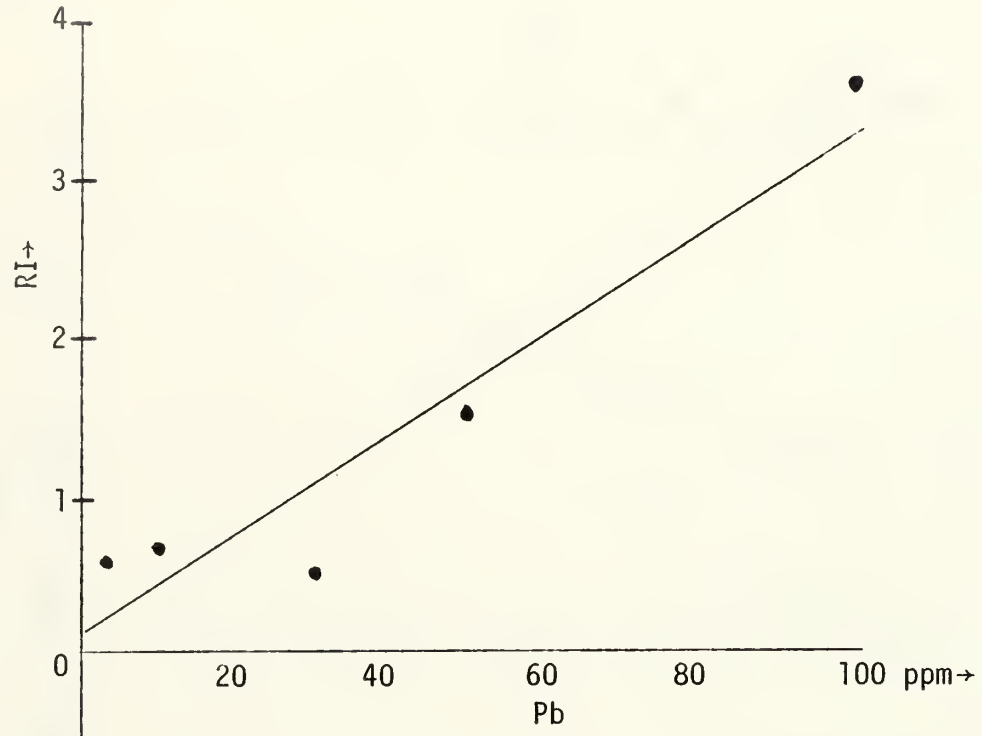
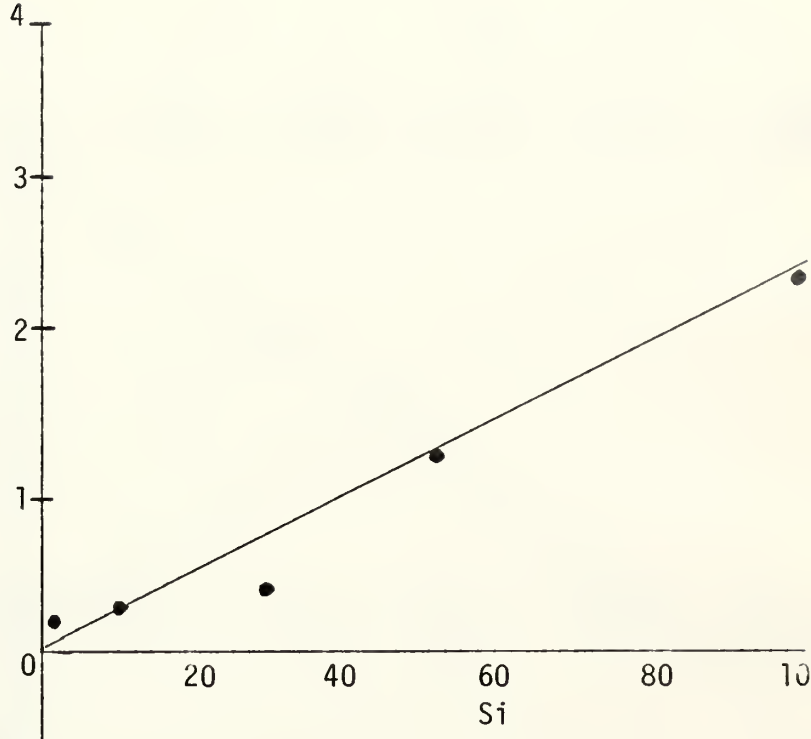
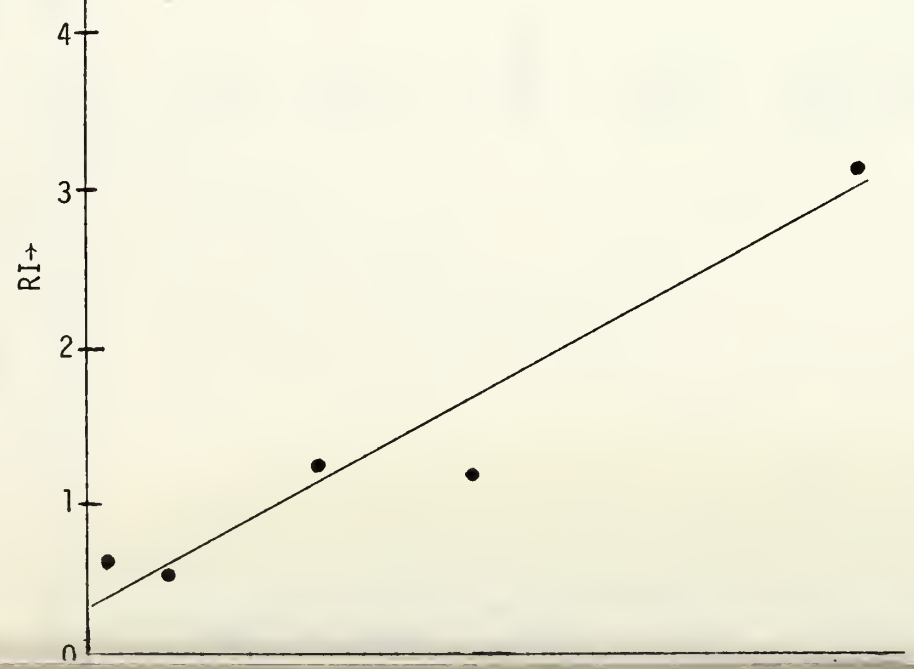


Figure 2 Continued







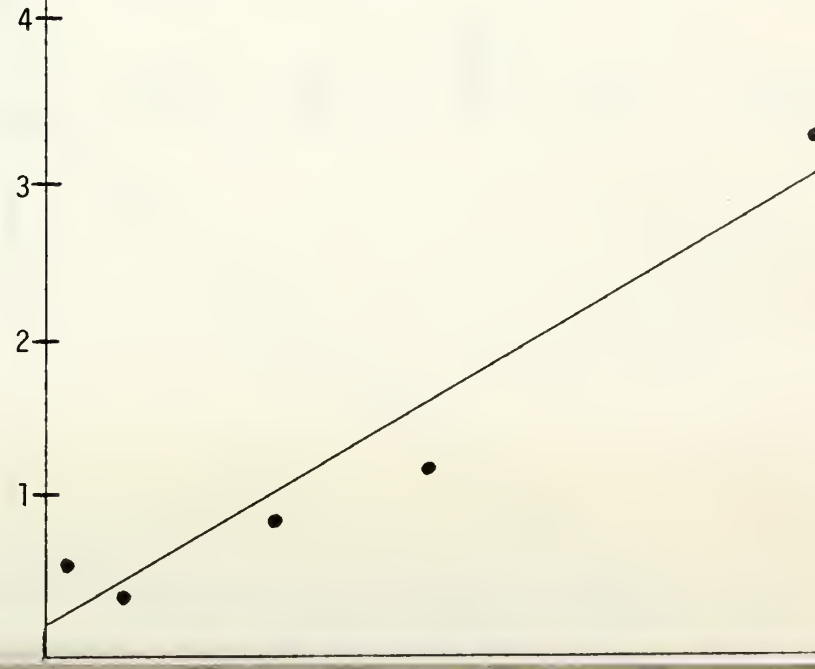


Figure 2 Continued

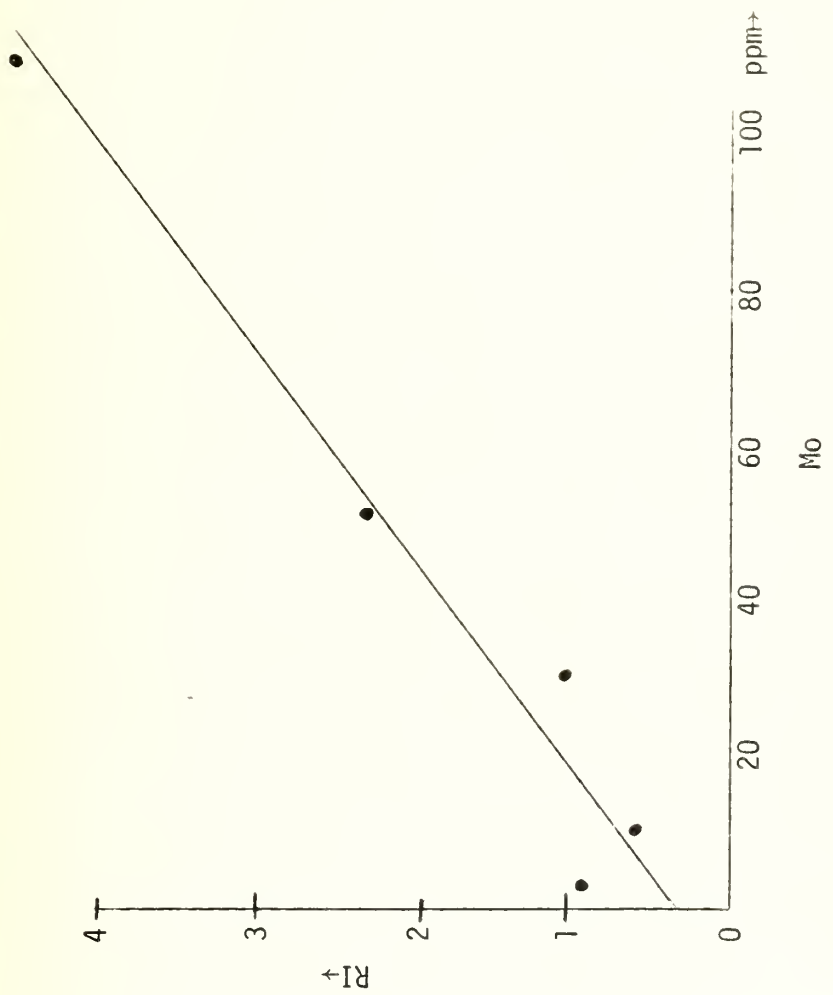


Figure 2 Continued

Linear fit Estimates	Fe	Ag	Aℓ	Cr	Cu	Mg	Na	Ni	Pb	Si	Sn	Su	Ti	M_0	Baird Atomic tolerance (typical)
b	-.114	-.173	.137	.117	.010	-.158	.403	.105	.084	-.028	.307	.125	.303	1.06	.259
m	.035	.029	.029	.022	.023	.040	.018	.034	.033	.023	.027	.031	.041	.040	.047
r	.983	.990	.958	.977	.961	.988	.914	.989	.957	.982	.962	.960	.975	.995	.999
Quadratic fit Estimates															
b	.412	.268	.596	.517	.38	.377	.631	.573	.582	.379	.684	.597	.745	1.090	.724
m	.015	.015	.012	.010	.012	.017	.009	.014	.013	.012	.011	.012	.014	.012	.016
r	.989	.997	.977	.990	.951	.987	.846	.993	.956	.983	.970	.972	.969	.979	.994

TABLE 3. Estimates of slope(m) and intercept(b) and correlation(r) for two models:

$RI = mc_0 + b$ (at top of table) and $\sqrt{RI} = mc_0 + b$ (bottom of table).

more indicate satisfactory fit; values in excess of .98 indicate quite close fit.

Values of RI computed for initial concentrations greater than 100 ppm, which were diluted to 100 ppm for analysis, were not significantly different from those for undiluted 100 ppm initial concentrations. Note: It was found that several sample runs for C_0 greater than 100 ppm had R-1 data identical with other sample runs. For example, data set 31 has the same R-1 data as data set 34 (and 30 has the same as 37). Thus, it appears the R-1 concentrations for runs with initial concentrations above 100 ppm (as shown on the computer print-outs) were in fact accomplished with undiluted R-1 standard at 100 ppm, in conjunction with other standards testing. If this is the case, no difference in RI due to dilution of more concentrated samples would exist, of course, for R-1 data.

Tolerances are needed for both accuracy and repeatability of sample runs. Following accepted statistical principles, the accuracy tolerances should depend on the inherent repeatability of the analysis process. Thus, with analysis procedures having high variance, one could detect only large differences in the standards under test (if one desired to control, at specified levels, the probabilities of committing errors in one's conclusions). Theoretically, in order to test whether two standards have the same concentration of a given element,

say iron, it is necessary to compare the difference in estimated levels in each standard, measured in standard deviation units, with a critical value taken from the statistical tables. For purposes of illustration, we describe such a procedure in what follows. If $X_1, X_2, \ldots, X_n$ denote analyses of iron made with the old standard, and $Y_1, \ldots, Y_n$ denote analyses of iron in the new standard (with analyses alternating between old and new, as is current (and good) practice), then

$$|T| = \frac{|\bar{X}-\bar{Y}|}{S_{\bar{X}-\bar{Y}}}$$

is compared with $t_{2n-2;1-\alpha/2}$, where

$\bar{X}$ is the average of n consecutive analyses with the old standard

$\bar{Y}$ is the average of n consecutive analyses with the new standard

$$S_{\bar{X}-\bar{Y}} = \left[\frac{\Sigma(X_i-\bar{X})^2 + \Sigma(Y_i-\bar{Y})^2}{n(n-1)}\right]^{1/2}$$ is the estimated standard deviation of $\bar{X}-\bar{Y}$, and

$t_{2n-2;1-\alpha/2}$ is the tabulated $(1-\alpha/2)100\underline{\text{th}}$ percentile of the t-distribution with $2n-2$ degrees of freedom.

A test would reject equivalence of the old and new standards (and thus would reject the new standard for iron content) at the α level of significance if $|T| > t_{2n-2;1-\alpha/2}$, that is, if $|\bar{X}-\bar{Y}| > S_{\bar{X}-\bar{Y}} \cdot t_{2n-2;1-\alpha/2}$.

The point of this illustration is not the test itself; rather, it is to demonstrate how a "tolerance," in this case $S \cdot t$, for testing <u>accuracy</u> $(\bar{X}-\bar{Y})$ is a linear function of the joint precision (repeatability), S. If different elements exhibit varying characteristics of change in repeatability with changes in initial concentration, then tolerance specifications should likewise vary over elements and initial concentrations. It is interesting to examine the accuracy and repeatability "acceptance" tolerances listed in Tables 4-14 and 4-15 of T.0.33A6-7-24-1 (enclosure 2 of our data from TSC, hereafter referred to as "Baird Atomic" acceptance tolerances) from this point of view. It is easily verified that, within each group of elements, AI_A is a linear function of RI_A. For example, for the group {Ni, Si, Aℓ, Be Cr}, $AI_k = 1.885RI_A + .233$, with a correlation very close to 1. (It is also interesting to note that RI_A in Table 4-15 of T.0. 33A6-7-24-1 is, within each element group, nearly linear in initial concentration, consistent with our finding that linear functions provide acceptable fits of the apparent relationships between RI and C_0.)

Comparison of the relationships estimated from T.0. 33A6-7-24-1 with the theoretical coefficients from the t-tables can provide some idea of the error rate levels one might achieve using the Baird Atomic Acceptance tolerances. Following the (2-sided, 2-sample t-test) argument above, theoretica

$$AI = \frac{\sqrt{2}\ t_{2n-2,1-\alpha/2}}{\sqrt{n}}\ RI \quad .$$

For example, with $n = 10$ analyses from each standard and $\alpha = .05$ (the probability of rejecting the new standard iron content, given it has in fact the same concentration of iron as the old standard), we would have

$$AI = \frac{\sqrt{2}\ (2.101)}{\sqrt{10}}\ RI = .940\ RI_A \quad .$$

From comparisons of Tables 4-14 and 4-15 of T.0.33A6-7-24-1, we find approximately (for all groups of elements) $AI_A \simeq 1.9(RI_A) + b$ where b is a "calibration error allowance" of about .25 ppm. In order to obtain the slope 1.9 in this relationship with the t-test with $n = 10$, one would need to take $\alpha \simeq .0005$. Based on this analysis, it appears that test procedures using the tolerances given in Table 4-14 give quite conservative tests; we suggest somewhat tighter tolerances with the procedure recommended below.

There appear to be two major goals in the standards testing activity. In roughly descending order of importance to the TSC, they are:

1) testing $R_1 = R_2$ for each element,
2) assuring analyses meet repeatability specifications for each element.

In addition to the statistical considerations, concerning setting of tolerances, discussed thus far (primarily the principle of setting tolerances in terms of repeatability attained by the analysis process), several operational considerations are involved. These can be stated in terms of the practical consequences of committing "type I" and "type II" errors in testing for each of the goals listed above. A type I error occurs whenever a satisfactory product (standard) is judged unsatisfactory by the test procedure. This usually occurs because data are obtained (by chance) that do not fairly represent the "typical" data produced by the procedure. A type II error occurs when a product that is actually unacceptable is judged acceptable by the test procedure.

General features of such procedures include:

1) Any screening or acceptance testing procedure will commit type I and type II errors from time to time, although the users of the procedure may not be aware of their occurrence,

2) as the type I error rate, α, is made smaller, the type II error rate, β, increases,

3) both α and β can be made smaller by increasing sample size, n, and

4) usually the type I error rate, α, together with n, are taken as the control variables; the value of β corresponding to a choice of α and n is thus determined.

From an operational point of view, α and n should be selected for each goal so as to give test procedures with error rates that reflect the importance of the goals and the seriousness (in terms of cost or loss) of committing type I and type II errors. For example, for the primary goal of testing $R_1 = R_2$, considerations include the implications of operating with a new standard having concentrations of one or more elements different from those of the previous standard, and the costs associated with rejecting a new batch of standard, even though it was acceptable. We realize that assessing such costs and losses may be impossible in practice, although even rough estimates can be useful in determining appropriate levels of α and n.

For establishing tolerance for accuracy-related tests ($R_1 = R_2$), the selection of α and n constitutes the tolerance. That is, in place of an absolute tolerance (such as "$\pm$ 3 ppm") we specify tolerances, relative to repeatability of the Analysis system, by setting α and n. This has the advantage of relating tolerances directly to the operating characteristics of the test procedure, with immediate operational interpretation. It should be noted that testing Accuracy is in reality testing relative accuracy. We are testing whether the new standard gives readings essentially the same as the old standard, not whether the new standard contains "3 ppm of Cu," for example. Because of the role of frequent recalibation of the spectrometers, the impossibility of maintaining absolute control of contaminant level in ppm is not a problem. Assuring that the relative contents of the old and new standards are essentially the same must (and will) suffice.

For establishing tolerances for testing precision, we also follow the principles discussed above. We have noted that, in absolute terms, the repeatability observed in sample runs will generally depend upon concentration levels, as well as the elements under test. Thus the repeatability tolerances must vary with concentration level and element. If good laboratory procedures are strictly adhered to a high value of RI would indicate spectrometer malfunction, rather than any defect in the standard being tested. Thus our suggested procedure includes monitoring the RI values,

but if RI is "too high" for some set of analyses, it is the operating procedure or the spectrometer which is suspect, not that the standard being tested was incorrectly prepared.

In the absence of clear notions concerning costs and losses due to commission of errors in testing for the various goals, we use "default values" of α and take $n = 10$ in the procedures we describe in the following section. After some experience with these procedures has been gained, these values can be adjusted if necessary to give rejection rates which suit the TSC.

II.4. The Test Procedure

Now let us describe the suggested procedure for acceptance testing of prepared reference standards. We shall call the prepared standard to be tested the candidate reference standard. Five different concentration levels (3, 10, 30, 50 and 100 ppm) are to be tested. As already mentioned, we recommend that the elements be analyzed individually, for each concentration, even though the spectrometer readings for all 13 (or 20) elements are determined simultaneously. If a candidate reference standard fails the test described in some one or more elements, at a given concentration level, the candidate must then be remixed, to bring the errant element(s) into line (if possible) and then retested for all elements, not just the one (or more) which originally failed.

Should the candidate fail a second time, it must then be discarded, or possibly remixed again for consideration as being acceptable at some higher or lower concentration level.

It is assumed that the spectrometer has been accurately standardized at 0 ppm and at 100 ppm, using a previously accepted primary reference standard. $n = 10$ burns are made of the candidate standard at each specified concentration level. Let $X_1, X_2, \ldots, X_{10}$ be the 10 readings gotten for a specified element and let $\bar{X}$ be their average, and RI the repeatability index for these 10, computed in the usual way. As a first step the RI value should be compared with the appropriate entry in Table 4. (See the discussion at the end of this section regarding the origin of Table 4.) If RI exceeds the tabled value, for the specified concentration-element combination, then the procedure or the spectrometer itself would appear to be faulty. The spectrometer should be re-standardized and a new set of 10 burns run, carefully following accepted laboratory procedures. If again RI, for the same element, is too large it would appear that the spectrometer is out of order; no further testing of the candidate reference standards can be accomplished until it is repaired.

Granted the RI value does not exceed the appropriate value in Table 4, a 99% (or some other level if more appropriate) confidence interval for the mean of the population from which the 10 numbers were selected is computed as follows (the values in Table 4 were computed from repeated runs made under ideal conditions. The values presented for RI in this table may in some cases be unrealistically low for daily use):

TABLE 4. Suggested Limiting Values for RI.

Element	3	10	30	50	100
Fe	.42	.54	1.33	2.27	5.04
Ag	.17	.49	1.33	2.13	5.31
Al	.73	.93	1.68	1.85	4.58
Cr	.46	.60	1.44	1.65	3.42
Cu	.25	.53	1.52	1.67	4.08
Mg	.30	.83	1.65	2.71	5.91
Na	.22	.94	1.82	2.05	4.74
Ni	.68	1.08	1.74	2.89	5.76
Pb	.88	.89	1.24	2.71	4.65
Si	.37	.60	1.46	2.00	3.48
Sn	1.07	1.38	1.57	1.75	4.48
Ti	.84	.94	1.55	2.99	4.60
Mo	1.00	1.00	1.92	3.32	7.53

the $99.5^{\underline{th}}$ quantile of the t-distribution with 9 degrees of freedom is $t_{.995} = 3.250$. The 99% confidence interval for the population mean then has endpoints $\bar{X} - (3.250)RI/\sqrt{10}$ and $\bar{X} + (3.250)RI/\sqrt{10}$, where RI is the repeatability index. [The general form for this $100(1-\gamma)\%$ interval is $\bar{X} \pm t_{1-\gamma/2}\, RI/\sqrt{n}$ where $t_{1-\gamma/2}$ is the $100(1-\gamma/2)^{\underline{th}}$ quantile from the t-distribution with $n-1$ degrees of freedom and n is the sample size, in case it is desired to change either the sample size or the confidence coefficient.] If the desired true concentration of the candidate standard is covered by the confidence interval, accept the candidate standard as having the correct concentration of the element analyzed. If the confidence interval does not cover the desired true concentration then it may not have the correct concentration. To verify this conclusion an additional 10 burns of the candidate standard should be made, alternating with burns of the primary reference standard of the same nominal concentration: candidate-primary-candidate-primary, etc. Let $Y_1, Y_2, \ldots, Y_{10}$ be the 10 new candidate readings with average $\bar{Y}$ and repeatability index RI_Y and let $Z_1, Z_2, \ldots, Z_{10}$ be the 10 primary standard values with mean $\bar{Z}$ and repeatability index RI_Z. Both RI_Y and RI_Z should be no larger than the appropriate entry in Table 4; follow the instructions above about repeating the burns if either of them exceed the tabular value. If both satisfy this requirement compute the joint repeatability index by

$$S = [\tfrac{1}{2}\,(RI_Y^2 + RI_Z^2)]^{1/2} .$$

This in turn can be used to compute a confidence interval for the difference in true concentration of the candidate and reference standards as follows: The 99.5th quantile of the t-distribution with 18 degrees of freedom is $t_{.995} = 2.878$. The 99% confidence interval for the difference in true concentrations then has end-points $\bar{Y} - \bar{Z} - (2.878)S/\sqrt{5}$ and $\bar{Y} - \bar{Z} + (2.878)S/\sqrt{5}$. If this interval contains zero accept the candidate standard and, if not, reject the candidate reference standard and conclude its true concentration is not the desired level. It then must be remixed or discarded as described above.

NOTE: It is possible that statistical significance and chemical significance are not identical and this procedure may prove too stringent (the criteria may be impossible to meet). That is, in chemical terms perhaps a 30 ppm standard could actually have a true concentration anywhere between 29 and 31 ppm, say, without causing any difficulties. Thus a candidate standard should be acceptable in this case if its true concentration is as low as 29 or as high as 31 ppm. In the procedure just described, then, the candidate standard should be initially accepted if 29 or 31 or any value in between is included in the confidence interval for its true concentration level. (In more general terms, accept the 30 ppm candidate if $30 + \Delta$ or $30 - \Delta$ or any number in between is covered by the confidence interval where Δ defines the limits of chemical significance.) If the 30 ppm candidate is initially rejected, and 10 more burns are alternated with the

30 ppm reference standard, accept the 30 ppm candidate if the confidence interval for the mean difference in the two concentrations includes -2 or 2 or any number in between. (Again if $30 - \Delta$ and $30 + \Delta$ define the limits of chemical significance, accept the 30 ppm candidate if -2Δ or 2Δ or any number in between is covered by the confidence interval for the difference.) With these modifications for chemical significance, the procedure described should prove a practical and useful way to control the quality of newly prepared standards.

Origin of Table 4.

The numbers in Table 4 were computed from data sets supplied by the TSC as an enclosure to their letter dated July 28, 1977. Data sets 1 through 9 contain 3 collections of 10 burns of primary reference standard R-1, by the Pensacola laboratory. RI was computed for each of these, for each element, giving 3 RI values for each element-concentration combination. These 3 RI's were pooled within each concentration-element combination, using the formula

$$RI_p = \sqrt{\frac{1}{3}\,(RI_1^2 + RI_2^2 + RI_3^2)} \quad .$$

In theory RI_p^2 is a constant times a χ^2-random variable with 27 degrees of freedom. If we let RI* be the repeatability index from 10 burns of a candidate standard (some specified element and concentration) the ratio $(RI^*)^2/RI_p^2$ has the F-distribution with

9 and 27 degrees of freedom and, with probability .99 this ratio should not exceed 3.16 or, equivalently, RI* should not exceed $RI_p\sqrt{3.16}$. This latter value is given in Table 4. Three entries in Table 4, Si-30, Sn-30 and Mo-3, did not seem reasonable when calculated from this formula, due to what appeared to be aberrant results in data sets 1 through 9. These have been adjusted slightl from what this formula would give. As indicated earlier, the numbers in Table 4 may be too conservative in some cases. In such situations larger limiting values for RI have to be chosen.

II.5. A Numerical Example

Assume the 10 readings gotten for a 30 ppm candidate standard are as given in Table 5. The average values, $\bar{X}$, and RI values are also listed there, as are the lower and upper 99% confidence limits computed from the formula discussed above. Note that none of the RI values exceed the appropriate entries in Table 4, so the next step is the computation of the confidence limits (given in Table 5). The confidence limits for Fe, Aℓ, Ni, Pb, and Si do include 30, the nominal level tested, so these elements appear to be at the correct concentration level. None of the confidence intervals for the remaining elements, however, contain 30 so they would all be suspect. Now let us suppose that chemical common sense dictates the true ppm content could be anywhere between 29 and 31 ($\Delta = 1$) and the candidate standard would be acceptable. This would mean that we want to see if 29 or 31 or any number in between is included between the confidence

limits for the remaining elements. With this change, Cr, Sn, Ti and Mo are now acceptable, but Ag, Cu, Mg and Na are still unacceptable. Thus, 10 more burns of the candidate, alternating with 10 burns of the 30 ppm primary reference standard are called for, with only the readings for Ag, Cu, Mg and Na to be analyzed.

Assume the values in Table 6 result. Again all RI values are acceptable (compared with entries in Table 4). Also given in Table 6 are the values for

$$S = \sqrt{\frac{1}{2}\,(RI_Y^2 + RI_Z^2)}$$

and the upper and lower confidence limits for the difference in mean concentration of the candidate and reference standards using the formula discussed above. Since each confidence interval includes zero we would conclude that the 30 ppm candidate is acceptable for all elements. (Granted that chemical common sense allows $\Delta = 1$, we would still have accepted the candidate if the 99% confidence limits for Na were, say, -3 and -1, since this interval includes -2.)

II.6. Summary of Calibration Standards Testing

a. Carefully standardize the spectrometer using the primary reference standard at 0 ppm and 100 ppm.

b. Following accepted laboratory techniques make 10 burns of the candidate standard at each prepared concentration: 3, 10, 30, 50 and 100 ppm.

c. For each element and concentration compute the average $\bar{X} = \frac{1}{10} \sum_{j=1}^{10} X_j$ and the repeatability index

$$RI = \sqrt{\frac{1}{9} \Sigma (X_i - \bar{X})^2} \ .$$

d. Compare RI for each element and concentration with the appropriate value in Table 4. If RI exceeds the value in Table 4 for any element-concentration combination, restandardize the spectrometer and carefully repeat 10 burns of the candidate at the same concentration and again compute RI for each element. If any RI exceeds the appropriate value in Table 4, the spectrometer should be checked before proceeding further. After the spectrometer is again in good working order, start again at a.

e. For each element-concentration combination compute the 99% confidence limits for true concentration: $\bar{X} - (3.250)RI/\sqrt{10}$, $\bar{X} + (3.250)RI/\sqrt{10}$. Let C_0 represent the nominal concentration level and $C_0 \pm \Delta$ the limits of chemical significance. If $C_0 - \Delta$, $C_0 + \Delta$ or any value in between lies between the confidence limits $\bar{X} \pm (3.250)RI/\sqrt{n}$, for <u>each</u> element-concentration combination, accept the candidate standard. If this is not true for some element-concentration combinations go to f.

f. For each concentration where $C_0 - \Delta$ and $C_0 + \Delta$ fall outside the confidence interval in e., repeat 10 burns of the candidate, alternating with burns of the primary reference standard of the same concentration. The following computations are made only for the elements, from e., whose true concentration is suspect. Let $\bar{Y}$, RI_Y be the average and repeatability index for the candidate and let $\bar{Z}$, RI_Z be the average and repeatability index for the primary reference standard at the same concentration, same element. Compute the 99% confidence limits for the difference in true concentration level for the two:

$$\bar{Y} - \bar{Z} - 2.878S/\sqrt{5} \ , \qquad \bar{Y} - \bar{Z} + 2.878S/\sqrt{5} \ ,$$

where

$$S = \sqrt{\frac{1}{2}\left(RI_Y^2 + RI_Z^2\right)} \ .$$

If -2Δ, 2Δ or any value in between lies between these confidence limits, that element appears to have an acceptable concentration level. If all element-concentration levels, which were suspect from e., satisfy this then conclude the candidate standard is acceptable at all concentrations tested. Any element-concentration for which this is not satisfied, appears to have an unacceptable concentration level and should be rejected.

TABLE 5. Candidate Readings

	Fe	Ag	Al	Cr	Cu	Mg	Na	Ni	Pb	Si	Sn	Ti	Mo
	29.1	32.5	30.6	31.4	33.0	32.7	32.5	30.5	29.0	28.2	30.3	30.6	30.3
	30.7	34.5	31.2	32.2	34.5	33.9	35.4	32.1	29.9	30.2	32.2	32.1	32.5
	29.8	33.5	29.3	31.0	33.2	32.5	34.9	30.2	29.7	30.5	31.4	30.1	32.1
	29.6	32.8	29.3	31.0	33.0	32.7	33.0	29.6	29.0	29.9	31.6	31.6	31.5
	29.5	33.7	29.1	30.9	33.4	32.7	33.9	29.6	28.4	29.9	29.5	30.1	30.5
	31.1	34.1	30.0	31.2	33.5	34.5	33.2	30.4	30.2	30.6	32.3	32.4	33.9
	29.7	33.3	30.0	31.3	32.9	32.9	33.7	30.5	29.3	29.8	31.7	31.6	32.4
	30.4	34.4	32.1	32.9	33.5	33.3	33.1	30.5	31.4	31.0	32.8	32.5	33.6
	29.5	33.1	29.7	30.3	32.9	32.8	32.3	30.2	29.6	29.9	31.0	30.5	32.9
	30.3	33.6	29.3	31.0	33.2	32.8	33.8	29.3	29.5	31.1	31.5	30.8	32.7
Averages	29.97	33.55	30.06	31.32	33.31	33.08	33.58	30.29	29.60	30.11	31.43	31.23	32.24
RI	.63	.66	.97	.73	.48	.64	.99	.77	.81	.83	.97	.92	1.19
Lower CI	29.32	32.87	29.06	30.57	32.82	32.42	32.56	29.50	28.77	29.27	30.43	30.28	30.98
Upper CL	30.62	34.23	31.06	32.07	33.80	33.74	34.60	31.08	30.43	30.95	32.43	32.18	33.42

TABLE 6. Candidate and Reference Readings

	Ag	Cu	Mg	Na
Candidate	33.3	32.4	32.7	33.9
	32.3	32.1	31.4	30.7
	34.3	33.3	34.4	34.7
	33.3	32.7	32.9	33.3
	34.8	33.2	33.6	34.5
	33.6	32.7	32.1	34.3
	33.3	32.4	32.9	32.7
	32.9	32.1	31.9	34.7
	33.6	32.4	33.8	33.4
	34.4	32.5	32.3	34.3
$\bar{Y}$	33.58	32.58	32.80	33.65
RI_Y	.75	.41	.93	1.23
Reference	33.5	33.3	33.5	33.9
	34.4	33.3	33.4	34.7
	34.6	33.2	33.6	34.1
	34.1	33.2	32.4	35.4
	34.9	33.1	32.6	34.8
	32.9	31.6	32.9	32.6
	31.5	31.1	30.0	33.4
	32.5	31.3	31.5	33.8
	33.9	32.2	32.3	35.0
	33.6	32.6	32.8	35.2
$\bar{Z}$	33.59	32.49	32.50	34.29
RI_Z	1.04	.88	1.08	.89
S	.91	.69	1.01	1.07
Lower CL	-1.18	- .80	-1.00	-2.02
Upper CL	1.16	.98	1.60	.74

III. LABORATORY CERTIFICATION

III.1. Introduction

Paragraph 2 of the project order MME-77-006 requires the development of statistical methodology to evaluate and certify the spectrometric laboratories participating in the joint oil analysis program. The evaluation of a laboratory is to be comprised of three sub-evaluations viz., an evaluation of the spectrometer performance, a comparison of the laboratory performance with that of another laboratory that is considered to have met certification criteria, and an assessment of the oil analysis evaluator's ability to make correct decisions based on the results of the analyses.

The methods we present in this paper are applicable for evaluating the spectrometric analyses results on a single element. As in the previous chapter, separate evaluations for the different elements are recommended and, of course, the same statistical methods are to be used with each element. The same is also true for different initial concentration levels in the standard oil samples; a separate statistical analysis for each initial concentration level is to be performed. The rest of the discussion, therefore will apply to the results of repeated independent

analyses (replications) on a single element with a fixed initial concentration level in the standard oil samples. However, a laboratory should be considered to have met all certification requirements only if it passes the statistical tests for each combination of element and concentration level.

The spectrometer evaluation methodology will require each laboratory to analyze a standard sample with a fixed initial concentration level, each day. If the spectrometer performance is to be examined at different concentration levels then daily analyses must be performed at each concentration level of interest. At the time a laboratory is due for certification, the data for the immediately preceding twelve months will be used.* The inter-laboratory comparison does not require any new data and all the required information can be extracted from the monthly correlation reports.

III.2. Spectrometer Certification

We propose a two-part procedure for determining if a spectrometer meets certification criteria. The first part is a macro test to see if during the preceding year, on the average, the accuracy and repeatability indices were within "acceptable limits." The acceptable limits we propose for usage are the maximum allowable accuracy and repeatability indices as given on page 8-2 of the JOAP Laboratory Manual of 1 May 1977. We recognize that these limits are quire conservative in the sense that they are not the tightest bounds possible. If a

*If the laboratory is new and has been in existence for less than one year, a modified procedure, described at the end of this section, may be used.

better set of bounds can be determined, perhaps based on past data, they should be used in the tests described herein. Part two is a micro test comprised of twelve separate analyses of the monthly results; this test is essentially a test for consistency.

Let X_{ij}, $i = 1,2,\ldots,12$; $j = 1,2,\ldots$ be the results of the spectrometric analyses for a specified combination of element and concentration level. The subscript i ranges over the twelve months and the subscript j represents the working days within each month. Thus, the total number of X's will be equal to the number of working days for the year. Let

n_i = number of data points for the i^{th} month

$N = \sum_{i=1}^{12} n_i$ = total number of observations

$\bar{X} = \sum_{i=1}^{12} \sum_{j=1}^{n_i} X_{ij}/N$ = average for the year

$S^2 = \sum_{i=1}^{12} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2/(N-1)$ = sample variance for the y

μ_0 = initial concentration level

A_0 = maximum allowable accuracy level

$R_0 = A_0/2$ = maximum allowable repeatability level

$\alpha = .05$ = significance level or Type 1 error probability

$Z_{.05} = -1.645$ = tabulated 5^{th} percentile of the standard normal distribution

$Z_{.975} = 1.96$ = tabulated 97.5^{th} percentile of the standard normal distribution

$\chi^2_{.05,N-1} = \frac{1}{2}\left[-1.645 + \sqrt{2N-3}\right]^2$ = approximate 5^{th} percentile of a chi-square distribution with N-1 degrees of freedom

$t_{.975,9} = 2.262$ = 97.5^{th} percentile of the student's t-distribution with 9 degrees of freedom

We assume that the X_{ij}'s are normally distributed with an unknown mean value μ and an unknown variance σ^2. Previous studies have shown that, as a general rule, the results of spectrometric analyses tend to be normally distributed.

a. Macro Test. This test consists of statistically establishing whether or not, the true accuracy index $|\mu-\mu_0|$ and the true repeatability index σ are below the maximum values A_0 and R_0, respectively. We first compute a 95% upper confidence bound for σ^2 as $[(N-1)S^2/\chi^2_{.05,N-1}]$ (that is,

$$P\left[\sigma^2 < \frac{(N-1)S^2}{\chi^2_{.05,N-1}}\right] = .95\right) .$$

Since it is required that $\sigma^2 < R_0^2$ we can conclude, with about 95% confidence, that the repeatability index is within acceptable

bounds provided that

$$\frac{(N-1)^2}{\chi^2_{.05}} < R_0^2$$

The chance that this procedure will result in a conclusion that the repeatability index is unacceptable, when in fact it is, is about 5%. Next, we obtain a 95% confidence interval for μ as $\bar{X} \pm Z_{.975}(S/\sqrt{N})$ (that is,

$$P\left[\bar{X} - Z_{.975}\frac{S}{\sqrt{N}} < \mu < \bar{X} + Z_{.975}\frac{S}{\sqrt{N}}\right] = .95\Big) \quad . \qquad (1)$$

The maximum acceptable accuracy index is A_0, which implies that $|\mu-\mu_0|$ must be less than A_0 or equivalently μ must satisfy the inequality constraint

$$\mu_0 - A_0 < \mu < \mu_0 + A_0 \qquad (2)$$

A combination of (1) and (2) will provide the criterion for acceptability of the accuracy index viz., conclude that the accuracy index for the spectrometer meets the certification criterion if

$$|\bar{X} - \mu_0| < A_0 - (1.96)\frac{S}{\sqrt{N}}$$

The probability of wrongly concluding that the accuracy index is unacceptable is about 5%. If both the accuracy index

and the repeatability index are found to be acceptable, the macro test has been met and we proceed to the next stage.

b. Micro Test. This is a procedure to check whether, on a monthly basis, the spectrometric analyses results are consistent and that there are no significant fluctuations from month to month. We do this by computing twelve 95% confidence intervals for the unknown mean μ, based on a sample of size 10 observations for each month. From among the n_i observations for the i^{th} month a sample of size 10 is selected; we suggest that every second observation starting with the second working day of each month be selected. As long as the spectrometric laboratories are not aware of the selection process it should not result in any systematic bias creeping in. It may happen that for certain months (February, for example) the selection scheme will not result in ten samples. If this is the case, additional samples to make up the difference should be taken at random from the remaining data for the month. Let $Y_{i1}, Y_{i2}, \ldots, Y_{i,10}$ be the ten measurements sampled for the i^{th} month and let

$$\bar{Y}_i = \sum_{j=1}^{10} Y_{ij}/10 \quad \text{be the sample mean}$$

and

$$S_i^2 = \sum_{j=1}^{10} (Y_{ij} - \bar{Y}_i)/9 \quad \text{the sample variance.}$$

The 95% confidence interval for μ for the i^{th} month will then be

$$\bar{Y}_i - (2.262)\frac{S_i}{\sqrt{10}} < \mu < \bar{Y}_i + (2.262)\frac{S_i}{\sqrt{10}}, \quad i = 1,2,\ldots,$$

As in the case of the macro test, we conclude that the accuracy index for the i^{th} month meets the certification criterion if

$$|\bar{Y}_i - \mu_0| < A_0 - (2.262)\frac{S_i}{\sqrt{10}}$$

This procedure will wrongly conclude that the results for a month do not meet certification criteria about 5% of the time. Now, let us examine the results of the "acceptance sampling" scheme for the twelve months in question. If the spectrometer performance is consistent throughout the year, the number of monthly acceptance sampling tests that will lead to a rejection, has a binomial distribution; the parameters of the distribution are $m = 12$ and $p = .05$. An examination of the binomial tables shows that about 98% of the time at least 10 monthly tests should result in acceptance. Thus, the micro test will conclude that the spectrometer does not meet the certification criterion if the number of "acceptance tests" that lead to acceptance is less than 10.

c. Examples. Annual laboratory certification is a new concept and will not be operational for a while. We will, therefore, use sample statistics derived from the validation data on standard samples (furnished by JOAP-TSC) for purposes of illustration of the methods described in this paper.

Macro Test:

Element: Cu

Initial concentration $\mu_0 = 100$ ppm

Max accuracy limit A_0: 10.5

Max repeatability limit R_0: 5.3

$N = 253$ = approximate number of working days in a year

$\bar{X} = 98.5$ = average of 253 spectrometer readings

$S = 3.84$ = sample standard deviation of 253 observations

$$\chi^2_{.05,252} = \frac{1}{2}\left[-1.645 + \sqrt{503}\right]^2 = 215.96$$

$$\frac{(N-1)S^2}{\chi^2_{.05,N-1}} = \frac{(252)(3.84)^2}{215.96} = 17.21$$

Since 17.21 is less than $R_0^2 = 28.09$ we conclude that the repeatability index meets the macro certification criterion.

$$|\bar{X} - \mu_0| = |98.5 - 100| = 1.5$$

$$A_0 - (1.96)S/\sqrt{N} = 10.5 - (1.96)(3.84)/\sqrt{253} = 10.03$$

Since $|\bar{X} - \mu_0| < A_0 - (1.96)S/\sqrt{N}$ the accuracy index also meets the macro certification criterion.

Micro Test:

The sample statistics and the results of the statistical analysis are presented in tabular form below:

Element: Cu; $\mu_0 = 100$; $A_0 = 10.5$

Month i	$\bar{Y}_i$	S_i	$\lvert\bar{Y}_i - \mu_0\rvert$	$A_0 - (2.262)\frac{S_i}{\sqrt{10}}$	Accept or Reject
1	98.1	1.37	1.9	9.52	Accept
2	96.2	2.57	3.8	8.66	Accept
3	100.7	2.31	0.7	8.85	Accept
4	101.4	2.37	1.4	8.80	Accept
5	101.8	3.19	1.8	8.22	Accept
6	99.7	3.16	0.3	8.24	Accept
7	100.7	2.95	0.7	8.40	Accept
8	99.2	4.26	0.8	7.45	Accept
9	97.0	2.00	3.0	9.07	Accept
10	100.7	2.41	0.7	8.78	Accept
11	98.3	1.57	1.7	9.38	Accept
12	97.6	2.46	2.4	8.74	Accept

Since each of the twelve monthly results is within acceptable limits the conclusion is that the spectrometer performance is consistent. It is apparent that with $A_0 = 10.5$ a monthly result will not be rejected unless the monthly average $\bar{Y}_i$

differs from μ_0 by a large amount; an examination of the validation data for standard samples shows that large differences occur very rarely, if at all. A more sensitive procedure would result if the maximum accuracy deviation is modified to $A_0' = A_0/2 = 10.5/2 = 5.25$. If this change is adopted the results of the macro test will be unaffected since $A_0' - (1.96)S/\sqrt{N} = 4.78$ and $|\bar{X} - \mu_0|$ is less than 4.78. For the micro test the monthly results for the second month will be unacceptable since $|\bar{Y}_2 - \mu_0| = 3.8$ is greater than $A_0' - (2.262)S_2/\sqrt{10} = 3.41$. However, because only one out of the twelve monthly tests leads to rejection the micro test would result in the conclusion that the spectrometer is consistent. Even if A_0 is changed to A_0', the maximum repeatability index $R_0 = A_0/2$ must be left unchanged since it is already a reasonably tight bound. It should be pointed out that in order to qualify for certification a laboratory has to pass each of the statistical tests for all combinations of elements and concentration levels for which data has been collected. With 20 elements and 5 concentration levels the number of combinations is 100. If $A_0' = A_0/2$ is used in place of A_0 itself, as the maximum accuracy limit, this will definitely increase the chance of at least one rejection out of the 100 combinations.

Some of the newer laboratories would have been in existence for less than a year. In these cases, full year's data will not be available and the tests will then have to be modified.

As an example, if data is available for six months or more both the macro and micro tests can still be performed. For the macro test the parameters N, $\chi^2_{.05,N-1}$ quoted earlier should be suitably modified. The parameter for the micro test will have to be replaced with the actual number of months for which data is available and a new "acceptance number" has to be determined from an examination of the tables of the binomial distribution. We recommend that the micro test not be used if the number of months is less than 6 since we believe that the test will not be very sensitive in this case.

III.3. Interlaboratory Comparison

As indicated in the introduction the laboratory certification scheme is to include a comparison of the performance of a laboratory that is to certified with that of another laboratory that has previously received certification. We believe that it is preferable to use a single laboratory such as the Pensacola laboratory as a standard against which all others are compared. The advantage of doing so is that the performance of the standard laboratory can be monitored on a regular basis to maintain a high performance level; besides, comparing all laboratories against a single standard laboratory

is a more equitable procedure. The comparison procedure will use data already available in the monthly correlation reports. At the time of certification, the results of the spectometric analyses of the standard samples of the preceding twelve months are extracted both for the laboratory in question as well as the Pensacola laboratory (the laboratories also analyze used oil samples under the correlation program but these are not of interest here). Let $X_1, X_2, \ldots, X_{12}$ be the spectrometer readings for the Pensacola laboratory and $Y_1, Y_2, \ldots, Y_{12}$ the corresponding readings for the laboratory to be certified. We will assume that

(i) $X_1, X_2, \ldots, X_{12}$ are independent and are normally distributed with means $\mu_1, \mu_2, \ldots, \mu_{12}$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_{12}^2$;

(ii) $Y_1, Y_2, \ldots, Y_{12}$ are independent and have normal distributions with means $\nu_1, \nu_2, \ldots, \nu_{12}$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_{12}^2$;

(iii) from past records (not including the twelve months data used for the comparison) for the Pensacola laboratory estimates $S_1^2, S_2^2, \ldots, S_{12}^2$ for the variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_{12}^2$ can be computed from samples of size $n = 10$ each.

The reasons for letting the μ's, ν's and σ^2's be different for different months is to allow for the possibility that the standard samples have different initial concentration levels and consequently non-identical means and variances. It is to be noted that the number of distinct μ's, ν's and σ^2's is equal to the number of distinct concentration levels in the correlation samples.

The implication of the assumption that both the X's and the Y's have the same variance within each month is that the emphasis in the interlaboratory comparison is on the accuracy and not so much on repeatability provided, of course, the repeatability indices are not too far apart.

With the above assumptions, the quantities

$$t_i = \frac{(X_i - Y_i) - (\mu_i - \nu_i)}{\sqrt{2}\, S_i}$$

are independent and each t_i has a student's t-distribution with $n-1 = 9$ degrees of freedom. If the performance of the laboratory to be certified is the same as that for the Pensacola laboratory, μ_i will be equal to ν_i. In this case, it can be shown that $P[|X_i - Y_i| > 2S_i] = .20$ approximately. In other words, if the means for the two laboratories are equal, the observed readings X_i, Y_i will differ by at least two standard deviations about 20% of the time. Now, consider the twelve

absolute differences $|X_i - Y_i|$, $i = 1,2,\ldots,12$. The number of times these differences will exceed twice the corresponding standard deviation S_i is a binomial random variable with parameters $m = 12$ and $p = .20$. From the binomial tables, it is observed that the number of differences that exceed twice the standard deviation will be less than or equal to five with probability .98; equivalently, the chance of observing six or more pairs that differ by more than two standard deviations is .02. This then provides a comparison test as summarized below:

Step 1: From past records for the Pensacola laboratory compute the sample variances $S_1^2, S_2^2, \ldots, S_{12}^2$ using a sample of size 10 for each computation. The number of different S_i^2 to be computed is equal to the number of distinct concentration levels used in the correlation samples. If all correlation samples have the same concentration level only one S^2 needs to be computed. From a practical point of view, the trimmed sample variances already available in the correlation reports may serve the purpose and may result in the saving of some labor. We believe that this change will not severely affect tne validity of the statistical procedure.

Step 2. Compute $|X_i - Y_i|$, $i = 1,2,\ldots, 12$.

Step 3: Let K = number of differences $|X_i - Y_i|$ that exceed $2S_i$.

Step 4: If $K \leq 5$ conclude that the laboratory under examination meets certification criteria.

Example: The data used in this example is fictitious although some of the numbers are sample statistics computed from the validation data for standard samples. Let X_i, Y_i, S_i and the initial concentration levels μ_{0i} (the concentration level in the standard sample for i^{th} month) be as in the table below.

Month	μ_{0i}	X_i	Y_i	S_i^*	$\lvert X_i - Y_i \rvert$	$2S_i$	Accept or Reject
1	3	2.88	2.86	.24	0.02	0.48	Accept
2	3	2.98	2.80	---	0.18	0.48	Accept
3	10	10.02	9.70	.44	0.32	0.88	Accept
4	10	9.71	9.32	---	0.39	0.88	Accept
5	30	29.84	29.01	1.45	0.83	2.90	Accept
6	30	29.73	28.48	---	0.25	2.90	Accept
7	50	50.45	50.14	1.93	0.31	3.86	Accept
8	50	50.79	49.89	---	0.90	3.86	Accept
9	100	102.0	102.1	4.52	0.10	9.04	Accept
10	100	101.3	105.4	---	4.10	9.04	Accept
11	100	102.1	100.1	---	2.00	9.04	Accept
12	100	102.0	98.2	---	3.80	9.04	Accept

*There are just five distinct concentration levels and hence only five different S_i.

There are zero rejections, so we conclude that the laboratory passes the comparison test.

The comparison test described above is applicable to most of the spectrometric laboratories participating in the Joint Oil Analysis Program. The requirement is that a laboratory is to have participated and analyzed standard samples under the correlation program for at least twelve months prior to the time the laboratory is due for certification. As indicated earlier the advantage is that no new data need be collected and the monthly correlation reports provide all the necessary information. Some of the newer laboratories, such as the Fort Riley laboratory, will not meet the requirement. We recommend that, in these cases, the following modified approach be adopted. JOAP-TSC will prepare twelve pairs of standard samples with a mixture of concentration levels; we suggest that the twelve pairs be comprised of two pairs each at 3, 10, 30 and 50 ppm and four pairs at 100 ppm concentration level. For each pair one sample will be analyzed at Pensacola and the other by the laboratory to be certified. The statistical analysis will be on the same lines as before, i.e. as given in Steps 1 to 4 above.

III.4. Evaluation Testing

The final subtask is the design of a test to be administered to the evaluators that are assigned to the spectrometric laboratories. The JOAP Laboratory Manual dated 1 May 1977 provides decision making guidance tables to aid the evaluator in his decision making process. Separate tables are provided for each type of equipment and contain numerical criteria relating the oil sample wearmetal concentration to the expected health of a component of the equipment. The recommended decisions are based on comparisons of the results of a used oil sample with that of a previous sample. The types of decisions an evaluator can make are (i) not to take any action; (ii) call for a more frequent sampling schedule; (iii) call for an immediate additional sample; (iv) recommend a maintenance action. The losses resulting from incorrect decisions by the evaluator can be quite high. A JOAP failure, i.e., an equipment that is being monitored by JOAP fails prior to detectio by JOAP can result in a loss of the equipment. Similarly, a JOAP miss, i.e., a JOAP recommended maintenance action which finds no discrepancies can be expensive. It is, therefore, very important that an evaluator be quite conversant with the basic facts about wearmetal concentrations and also have sufficient experience with analyzing sample results to look for trends and shortrun features such as a sudden rise in concentration levels right after overhaul. We suggest that the examination be in

two parts. The first part consists mostly of multiple choice questions which will test the basic knowledge about wearmetal concentrations that is critical for the various types of equipment being monitored. The second part will present actual historical data to illustrate the kinds of trends and the ambiguities that an evaluator will encounter. The test will examine his performance as gauged by the number of correct decisions made.

A set of sample questions testing basic knowledge are presented below.

(1) Spectrometric analysis will not detect

a) worn, misaligned or scored gears

b) broken piston rings and bands

c) failures due to fluid starvation

d) loose or defective valve guides

e) chips or wearmetal particles visible to the eye

(2) Explain in two or three sentences the effect of each of the following on the integrity of spectrometric analyses

a) contamination

b) electrodes

c) calibration standards

d) electrolytic corrosion

e) new or recently overhauled components

(3) Briefly describe the six steps to be followed in evaluating the sample results of incoming oil samples.

(4) If for aircraft types T-1A, T-33A, T-33B or QT33A, a sudden increase in Fe and Mg is observed the recommended action is to inspect

a) accessory drive assembly oil pump

b) main starter housing assembly

c) main bearing seals

(5) For F-101/F-102 aircraft the most significant and critical wearmetal is

a) Fe

b) Mg

c) Cu

d) Ag

e) Cr

(6) For F-84, B-57 aircraft the most significant wearmetal is

a) Fe

b) Mg

c) Cu

d) Ag

e) Cr

The above questions are based on information contained in the JOAP Laboratory Manual. For questions (4), (5) and (6) appropriate cutaways of the equipment may be provided. The equipment types selected to base the questions for the Navy evaluators should be Navy aircraft and helicopters; similarly for the other services.

We recommend that in the second part of the examination case histories illustrating the following situations be presented

a) Slow and steady increase in wearmental concentration but there is no potential failure
b) slow and steady increase in concentration level but the level has passed a critical stage
c) sample results after a recent overhaul showing a sudden increase in a wearmetal concentration
d) a JOAP failure
e) a JOAP hit
f) one or more ambiguous or marginal situations where either a maintenance action or no action would be considered reasonable
g) a case where there is a build up in Fe concentration due to corrosion.

IV. GRAPHITE ELECTRODES

IV.1. Introduction

The accuracy of readings produced by a batch of electrodes is of primary importance in judging the acceptability of the batch for use in the oil analysis program. The repeatability characteristics of the electrodes are also of some importance in judging acceptability. If a batch of electrodes scores badly on repeatability one can expect a number of spurious readings, including ones which may be too low (possibly missing a significant increase in some contaminant in a used oil sample) and ones which may be too high (possibly indicating a high contaminant reading when the level has not changed). Thus it is suggested that both repeatability and accuracy be considered in judging the acceptability of a new batch of electrodes.

The judgments of whether the new batch of electrodes is acceptable with respect to accuracy and repeatability can bes be made by comparison with readings gotten, on the same prepared oil sample, by using electrodes from a previously accepted batch. It is suggested that the elements of interest be considered one after another. For convenience it is assumed that a 10 ppm primary reference standard is used. A different oil standard could be used if desired.

IV.2. Acceptance Criteria for Graphite Electrodes

The suggested procedure calls for analyzing the spectrometer readouts one element at a time, to ensure that the electrodes are uncontaminated by any element of interest. To distinguish between the readings gotten with the new batch of electrodes versus those from the previously accepted batch we shall use a double subscript, the first subscript equalling one if the reading is made with an electrode from the new batch and this first subscript equals two if the reading is made with an electrode from a previously accepted batch. The second subscript distinguishes between the several readings made with the same type of electrode. We shall assume n_1 samples are analyzed with the new electrodes and n_2 with the old. (There is no special reason that we would have $n_1 \neq n_2$; the formulas presented allow for either $n_1 = n_2$ or $n_1 \neq n_2$.)

Thus the element readings from the new batch are $X_{11}, X_{12}, \cdots, X_{1n_1}$ and from the previously accepted batch they are $X_{21}, X_{22}, \cdots, X_{2n_2}$. For each set of readings we can compute the sample means:

$$\text{new batch} \qquad \bar{X}_1 = \frac{1}{n_1}(X_{11} + X_{12} + \cdots + X_{1n_1})$$

$$\text{previously accepted} \qquad \bar{X}_2 = \frac{1}{n_2}(X_{21} + X_{22} + \cdots + X_{2n_2})$$

and the repeatability indices:

new batch $$s_1 = \sqrt{\frac{(X_{11}-\bar{X}_1)^2+(X_{12}-\bar{X}_1)^2+\cdots+(X_{1n_1}-}{n_1 - 1}}$$

previously accepted $$s_2 = \sqrt{\frac{(X_{21}-\bar{X}_2)^2+(X_{22}-\bar{X}_2)^2+\cdots+(X_{2n_2}-\bar{X}}{n_2 - 1}}$$

The comparison of the two sets of readings is done in 2 steps. First we shall test the hypothesis that the repeatability index for the new batch does not exceed the index for the old. Granted this is accepted, we then will test the hypothesis that the mean reading for the new batch does not exceed the mean reading for the old.

To test that the new repeatability index does not exceed the old we compute s_1^2/s_2^2 and compare this ratio with a value from an F table with n_1-1 and n_2-1 degrees of freedom. Which entry to use is determined by the value desired for the probability of rejecting the new batch because of bad repeatability, when in fact it has an acceptable repeatability index. Suppose we set this probability at .01 and denote the tabular entry by $F_{.99}$. We then conclude the new batch is acceptable with respect to repeatability if $s_1^2/s_2^2 \leq F_{.99}$; otherwise we conclude it is not.

Granted that we find $s_1^2/s_2^2 \leq F_{.99}$, we then proceed to test the equality of mean readings. We first compute the combined repeatability index (pooled standard deviation) by

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

We then compute the test statistic

$$\frac{|\bar{x}_1 - \bar{x}_2|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which is compared with an entry from the t-distribution table. Again the entry to use is determined by the probability desired of concluding the new batch is not acceptable in accuracy, when in fact, it is acceptable. Suppose we set this probability at .01; we need the quantile $t_{.995}$ from the t-distribution with $(n_1 + n_2 - 2)$-degrees of freedom. We then say the batch is acceptable with respect to accuracy if

$$\frac{|\bar{x}_1 - \bar{x}_2|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{.995} ;$$

otherwise we reject the batch because of poor accuracy.

As described, this test is "two-tailed" and the new batch of electrodes would be declared unacceptable if $\bar{X}_1 - \bar{X}_2$ gets too large either positively or negatively. A large positive difference may be rightly attributed to possible contamination of the new batch of electrodes. A large negative difference, however, would seem to indicate that the previously accepted batch of electrodes contains a higher concentration of the element being analyzed than does the new batch. Logically one would not want to reject the new batch in this case. If this case occurs for one or more elements the procedure followed should be closely examined and the possibility of contamination of the old batch should be investigated.

This procedure is illustrated numerically below, assuming $n_1 = n_2 = 15$ samples analyzed with both the new and old electrodes. Although they are not written in that order, it is assumed that the analyses with the old and new electrodes are done alternately, to protect against a possible drift of the spectrometer during the period of analysis. The sample sizes of $n_1 = n_2 = 15$ are used for illustration only. In acceptance testing of large batches of material MIL STD 105D should be consulted regarding appropriate sample sizes. The assumed readings (for 10 ppm standard) are

X_{1j}	X_{2j}
9.5	9.5
10.1	8.8
9.8	9.1
9.4	8.9
9.6	9.2
9.6	9.3
9.5	9.5
10.1	9.4
9.7	9.4
9.7	10.1
10.0	9.3
10.2	9.0
10.0	9.8
10.0	9.5
9.7	9.4

We find $\bar{X}_1 = 9.79$, $\bar{X}_2 = 9.35$, $s_1 = .255$, $s_2 = .333$, and thus

$$\frac{s_1^2}{s_2^2} = .58 \ ;$$

Since $F_{.99}$ is about 3.5, with $n_1-1 = 14$ and $n_2-1 = 14$ degrees of freedom, we would accept the new batch for repeatability. We then compute

$$s_p = \sqrt{\frac{14(.255)^2 + 14(.333)^2}{28}} = .297$$

and

$$\frac{|\bar{x}_1 - \bar{x}_2|}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{|9.79 - 9.35|}{.297 \sqrt{\frac{2}{15}}} = 4.06 \ .$$

We find, with $n_1 + n_2 - 2 = 28$ degrees of freedom, $t_{.995} = 2.763$, and, since

$$4.06 > 2.763$$

we would reject the new batch in terms of accuracy. From these sample results it would appear that the new batch, on the average, gives a reading .44 ppm higher than that obtained with electrodes of the old batch, for this element. It may well be that such a difference is not practically significant, especially when one considers the acceptable equipment accuracy and repeatability indices in Tables 4-14 and 4-15, pages 4-55, 4-56 of T.O. 33A6-7-24. These tables give the acceptable accuracy index, for 10 ppm iron concentration, to be 2.21 ppm and the acceptable repeatability index (based on $n = 10$ analyses) to be .94 ppm. Since the accuracy index is the absolute value of the difference between a sample average reading and the assumed true concentration in the oil, this would imply an acceptable difference in two sample

averages of $2(2.21) = 4.42$ ppm. The acceptable pooled standard deviation for two samples of size 10 then would be

$$s_p = \sqrt{\frac{9(.94)^2 + 9(.94)^2}{18}} = .94$$

and the implied acceptable value of the t-statistic would be

$$\frac{4.42}{.94\ \sqrt{2/10}} = 10.51$$

With 18 degrees of freedom, a random variable with the t-distribution will exceed 8.115 with probability 10^{-7}. The implied acceptable t value of 10.51 above would occur with probability considerably less than 10^{-7}. This means that, if the two electrode batches are uncontaminated, there is less than 1 chance in 10 million of the t-statistic being this large. Therefore the tabled values mentioned do not seem to provide reasonable values for deciding the acceptability of electrode batches. Even if one allows the difference in mean readings of two samples of size 10 to be only 2.21, the table accuracy index value, this still implies an acceptable t-value of

$$\frac{2.21}{.94\ \sqrt{2/10}} = 5.26$$

which has probability of about .00005 of occurring if both batches are uncontaminated. If the new batch is contaminated, and the old is not, this magnitude for the t-statistic is much more likely to be observed. Thus values of t this extreme should not be called acceptable, because of the size of the associated large probability of accepting a contaminated batch.

It still may be desirable to allow, say c ppm difference in apparent content of the contaminant before rejecting the new batch. This may be accomplished as follows: If $\bar{X}_1 > \bar{X}_2$ accept the new batch unless

$$\frac{\bar{X}_1 - \bar{X}_2 - c}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{.995}$$

and if $\bar{X}_1 < \bar{X}_2$ accept the new batch unless

$$\frac{\bar{X}_1 - \bar{X}_2 + c}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -t_{.995}$$

With the above values $\bar{X}_1 = 9.79$, $\bar{X}_2 = 9.35$, $s_p = .297$, $n_1 = n_2 = 15$ and with $c = 1$, we have $\bar{X}_1 > \bar{X}_2$ and thus we compute

$$\frac{9.79 - 9.35 - 1}{.297\ \sqrt{2/15}} = -5.16$$

which is smaller than 2.763 so the new batch would be accepted.

There are two possible errors which could be made in considering a new batch of electrodes: A contaminated batch may be accepted (called Type II error) or a good batch may be rejected (called Type I error). For any two specific sample sizes n_1 and n_2 the smaller that one makes the probability of type I error the larger the probability of the type II error, and vice versa. Because of this one may not want to use such extremely small probabilities of type I error as would be suggested by the values in Tables 4-14 and 4-15 of T.O. 33A6-7-24-1 mentioned earlier.
Sample sizes of at least $n_1 = n_2 = 30$ and the probability of rejecting a good batch set at .01, for both the F and t statistics used, should provide a useful acceptance criteria. Mil STD 105D should be consulted for reasonable sample sizes in acceptance sampling of large batches of material.

This 2-stage test, or its adaptation, should be carried out in turn for each element of interest. If the new batch is rejected for any one or more elements, these electrodes should be declared unacceptable.

IV.3. Summary of Acceptance Criteria

Samples of n_1 and n_2 electrodes are selected from the new and old batches, respectively. Each electrode is used only once. The instrument should be accurately calibrated using electrodes of the old batch, with an accurately prepared oil standard. The burns with new and old electrodes are done alternately: new, old, new, old, etc. For each element of interest the acceptance procedure is

(1) Compute the average reading for the new batch

$$\bar{X}_1 = \frac{1}{n_1} (X_{11} + X_{12} + \cdots + X_{1n_1})$$

(2) Compute the average reading for the old batch

$$\bar{X}_2 = \frac{1}{n_2} (X_{21} + X_{22} + \cdots + X_{2n_2})$$

(3) Compute the repeatability index for the new batch

$$s_1 = \sqrt{\frac{(X_{11}-\bar{X}_1)^2 + (X_{12}-\bar{X}_1)^2 + \cdots + (X_{1n_1}-\bar{X}_1)^2}{n_1 - 1}}$$

(4) Compute the repeatability index for the old batch

$$s_2 = \sqrt{\frac{(X_{21}-\bar{X}_2) + (X_{22}-\bar{X}_2)^2 + \cdots + (X_{2n_2}-\bar{X}_2)^2}{n_2-1}}$$

(5) Compute s_1^2/s_2^2 and compare with $F_{.99}$, n_1-1, n_2-1 degrees of freedom. These may be found in "Tables of Common Probability Distributions," P.W. Zehna, D.R. Barr, Naval Postgraduate School Technical Report NPS 55ZeBn 0091A, pages 16-21 or some equivalent source. Use column $n = n_1$-1 (interpolate if necessary), major row $m = n_2$-1, minor row label .99. If $s_1^2/s_2^2 \geq F_{.99}$, reject the new batch for poor repeatability. If $s_1^2/s_2^2 < F_{.99}$, go on to 6.

(6) Compute the combined repeatability index

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

(7) Compute

$$\frac{|\bar{X}_1 - \bar{X}_2|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

(8) Find $t_{.995}$ from column .995, row $n = n_1 + n_2 - 2$, page 23, in "Tables of Common Probability Distributions," P.W. Zehna, D.R. Barr, Naval Postgraduate School Technical Report NPS 55Ze Bn 0091A, or some equivalent source. If $n_1 + n_2 - 2 > 30$, use $t_{.995} = 2.575$.

(9) If

$$\frac{|\bar{X}_1 - \bar{X}_2|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{.995}$$

the readings are acceptable for this element. Go on to analyze another element, starting at 1.

(10) If

$$\frac{|\bar{X}_1 - \bar{X}_2|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \geq t_{.995}$$

the performance of the new batch of electrodes is unacceptable.

IV.4. A Statistical Test to Evaluate Trace Metal Content of Graphite Electrodes as Determined on the A/E 35U-3 Spectrometer.

Just as with the acceptance criteria described above, the evaluation of the trace metal content of the new graphite electrodes is most appropriate measured relative to readings gotten with electrodes of known quality. The procedure for accomplishing this is described below.

Assume n_1 burns of the selected reference standard (say, 10 ppm) have been made with electrodes from the new batch. The discussion is pertinent for each element in turn and we let $X_{11}, X_{12}, \ldots, X_{1n_1}$ be the spectrometer readings for iron, say, and let $\bar{X}_1$ and s_1 be the average and repeatability index, respectively, for these n_1. Let $X_{21}, X_{22}, \ldots, X_{2n_2}$ be the spectrometer readings for this same oil standard using electrodes from a previously accepted batch. The average reading using the previously accepted electrodes then is $\bar{X}_2$ and their repeatability index (standard deviation) is s_2. A good measure of the excess iron trace metal content in electrodes of the new batch versus those previously accepted, is given by $\bar{X}_1 - \bar{X}_2$. It is easy to compute an interval with the property that we know how likely it is that the true average excess of the iron reading (new batch versus old) is included in the interval. This again requires values from the t-distribution and requires the pooled standard deviation (repeatability index):

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

If we want an interval which we are $100\gamma\%$ sure includes the true excess, we need $t^* = t_\gamma$ from the t-distribution

with $\nu = n_1 + n_2 - 2$ (row ν, column γ). Then we can be $100\gamma\%$ sure the true average excess does not exceed

$$\bar{X}_1 - \bar{X}_2 + t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad .$$

This is illustrated below.

Let us use the same data that was used in the acceptance criteria discussion above. Thus we have

$$n_1 = n_2 = 15, \quad \bar{X}_1 = 9.79, \quad \bar{X}_2 = 9.35, \quad s_p = .297$$

and we found $t_{.995} = 2.763$ with 28 degrees of freedom. Then we can be $100\gamma\% = 99.5\%$ sure the excess iron contaminant in the new batch, relative to the old, is no larger than

$$9.79 - 9.35 + 2.763(.297) \sqrt{\frac{2}{15}} = .74 \text{ ppm.}$$

IV.5. Variance Contributed by Electrode

To identify the variance contributed by the new batch of electrodes again let us discuss estimation on an element by element basis. We shall explicitly discuss the procedure and formulas for iron, say, with the understanding that the same procedure and formulas can be applied in turn for copper, aluminum, magnesium, etc.

We assume n_1 electrodes have been selected from the new batch, each to be used in analyzing a sample of the same oil, say a prepared standard containing 10 ppm of iron. Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ be the n_1 iron readings produced by the spectrometer using these electrodes from the new batch. We also assume we have n_2 electrodes from a previously accepted batch, each used to analyze a sample from the same oil standard. Denote these iron readings by $X_{21}, X_{22}, \ldots, X_{2n_2}$. The total variance, then of these $n_1 + n_2$ iron readings is a constant times the sum of the squares of each individual reading less the overall mean:

$$\sum_{i=1}^{2} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$$

where the overall mean is

$$\bar{X} = \frac{1}{n_1 + n_2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} X_{ij} .$$

This total sum of squares can be partitioned into two parts:

$$\sum_{i=1}^{2} \sum_{j=1}^{n_i} (X_{ij}-\bar{X})^2 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (X_{ij}-\bar{X}_i)^2 + \frac{n_1 n_2 (\bar{X}_1-\bar{X}_2)^2}{n_1 + n_2}$$

where

$$\bar{X}_1 = \frac{1}{n_1} \sum_{j} X_{1j} , \qquad \bar{X}_2 = \sum_{j} X_{2j}$$

are the average readings for the two electrodes. The first of these

$$\sum_{i=1}^{2} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 = \sum_{j=1}^{n_1} (X_{1j} - \bar{X}_1)^2 + \sum_{j=1}^{n_2} (X_{2j} - \bar{X}_2)^2$$

is just the sum of squares of the readings for each electrode about its own average value: part of the variability of the $n_1 + n_2$ readings is given by the variability within readings by the same electrode type. The remaining term

$$\frac{n_1 n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2)^2$$

is a constant times the square of the difference between the two averages: the remainder of the variability in the $n_1 + n_2$ readings is related to the difference in average readings of the two electrode types. This partition of the total sum of squares is frequently called an analysis of variance; it breaks the total variance into parts which can then be compared. The discussion of acceptance criteria in paragraph c above actually is using this same partition although it is not described in that way.

Isolation of the variance due to the electrode type may be done in a relative sense as follows: Let us assume that the variance of a reading, using an electrode from the new batch is

$$V[X_{1j}] = \sigma^2 + \sigma_1^2$$

where σ^2 is the variance due to the instrument, oil standard used, etc., and σ_1^2 is the contribution from the new electrode batch. Similarly, assume the variance of a reading from the previously accepted batch is

$$V[X_{2j}] = \sigma^2 + \sigma_2^2$$

where σ^2 is the same as before, since the same instrument, oil standard, etc., are used with these readings, and σ_2^2 is the contribution from the old electrode batch. It can be shown that

$$s_1^2 = \frac{1}{n_1 - 1} \Sigma (X_{ij} - \bar{X}_1)^2$$

is an unbiased estimate of $V[X_{1j}]$, i.e., of $\sigma^2 + \sigma_1^2$ and that

$$s_2^2 = \frac{1}{n_2 - 1} \Sigma (X_{2j} - \bar{X}_2)^2$$

is an unbiased estimate of $V[X_{2j}] = \sigma^2 + \sigma_2^2$. The difference, $s_1^2 - s_2^2$, then gives an unbiased estimate of $\sigma_1^2 - \sigma_2^2$, the differences in variance contributed by the two types of electrodes, since the term contributed by the instrument and standard cancels off in forming the difference. If, for example, we found $s_1^2 = .8$, $s_2^2 = .7$ then $s_1^2 - s_2^2 = .1$ is the estimated excess variance for the new electrode batch versus the previously accepted batch. Note that this measure is a function of the repeatability indices only and is unaffected by the accuracy indices of the two batches.

DISTRIBUTION LIST

	COPIES
Defense Documentation Center Cameron Station Alexandria, VA 22314	2
Library, Code 0142 Naval Postgraduate School Monterey, CA 93940	2
Office of Research Administration Code 012A Naval Postgraduate School Monterey, CA 93940	1
U. S. Army DARCOM Maintenace MGT Center/DRXMD-MS Lexington, KY 40511 Attn: G. Brown	1
JOAP-TSC NARF Code 360 Pensacola, FL 35208	3
NARF Code 440 Pensacola, FL 35208 Attn: R. Kight	1
SA-ALC/MMETP Kelly AFB, TX 78241 Attn: W. Walden	
SA-ALC/ACDCK Kelly AFB, TX 28241	1
Naval Postgraduate School Monterey, Ca. 93940 Code 55	
Attn: D. R. Barr	5
H. J. Larson	2
R. Stampfel	1
Code 53	
T. J. Jayachandran	3